OCR-software getest

Eenvoudig tekst herkennen

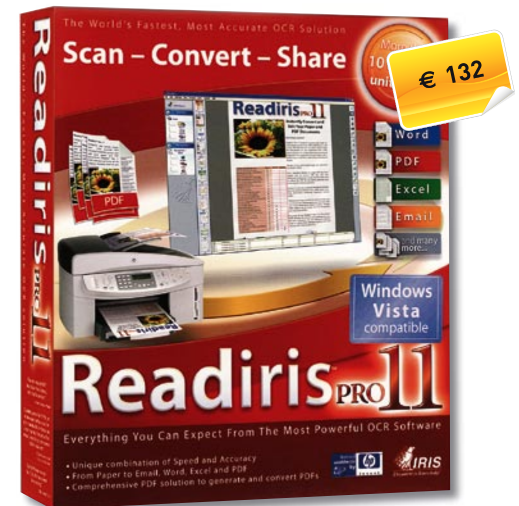
Vijf jaar geleden waren we al blij als we ingescande documenten met een nauwkeurigheid van 95% konden omzetten in tekst. Vandaag is er veel meer mogelijk, maar het aanbod is verschrompeld tot drie hoofdrolspelers, die we in deze test letterlijk tegen het licht houden. DIRK SCHOOFS



Dankzij OCR ('Optical Character Recognition' of optische tekstherkenning) kan je in een digitale afbeelding teksten en plaatjes van elkaar scheiden en afzonderlijk inlezen. Je scant bijvoorbeeld een tekst in, en even later heb je een Word-, Excel- of pdf-document op je bureaublad staan dat je inhoudelijk kan aanpassen. Het OCR-programma analyseert namelijk de afbeelding en gaat daarbij op zoek naar herkenbare tekens en zet die om in digitale tekst. Niet alleen het lettertype, maar ook de taal en soms zelfs de opmaak worden zo overgezet. OCR is met andere woorden het middel bij uitstek in de strijd tegen de papierberg.
De drie hoofdrolspelers gaan erg ver met hun producten. OmniPage van het Amerikaanse Nuance (het voormalige ScanSoft), FineReader uit de Russische ABBYY-stal en Readiris Pro van de Belgische beursgenoteerde I.R.I.S. Group verwerken zelfs met gemak lijvige pdf-documenten die je van het internet laadt. Alle drie herkennen ze tekst in meer dan 125 talen en bevatten ze snufjes die tekstherkenning tot op een hoog niveau tillen.

Readiris Pro 11.1

Nauwkeurige tekstherkenner



Van Readiris zijn er twee versies op de markt: de Corporate Edition en de Pro-versie. Die laatste richt zich tot de veeleisende thuisgebruiker en kleine bedrijven. De installatie is alvast een fluitje van een cent; na enkele minuten draait de Nederlandse versie vlekkeloos op ons systeem. Het programma herkent tekst in 126 talen. Zelfs de Griekse, Baltische en Cyrillische karaktersets vormen geen probleem, ook niet als er op één pagina verschillende talen voorkomen. In het begin gebruik je best de wizard, die je doorheen de verschillende stappen leidt. Ofwel open je een afbeelding die je vooraf hebt gescand, ofwel spreek je de scanner rechtstreeks aan vanuit het programma. De toepassing herkent afbeeldingen in jpg-, bmp- of tif-formaat. Erg handig is ook dat je met Readiris Pro een pdf-document van verschillende pagina's (maximum 50) in één enkele opdracht kan laten inlezen en omzetten in bijvoorbeeld Word of pdf. Wil je meer dan 50 pagina's door de herkenningsmolen draaien, dan ben je aangewezen op de Corporate-versie. Readiris Pro heeft trouwens nog meer in petto om meerdere documenten tegelijk te verwerken. Zo kan je een tijdsinterval instellen om tijdens het scannen van een boek pagina na pagina om te slaan en op de scanner te leggen.

Streepjescodes

Indrukwekkend is de mogelijkheid om een storende achtergrondkleur te onderdrukken, zodat de tekst op die pagina's zich veel nauwkeuriger laat analyseren. Je kan met Readiris Pro 11.1 zelfs streepjescodes omzetten. Naar eigen zeggen kan het ook handschrift herkennen, maar dat draaide in onze test iets anders uit. Zolang de letters niet aan elkaar geschreven zijn, lukt het nog wel, maar dat is bij de meeste handschriften natuurlijk niet het geval. Je kan het programma wel trainen, maar overtypen lijkt ons toch net iets sneller.
De nauwkeurigheid waarmee dit OCR-pakket tekst herkent, is indrukwekkend. Zelfs tabellen en opgemaakte tekst met daartussen foto's of andere illustraties, vormen geen enkel probleem. Readiris zorgde voor een quasi perfecte output in Word, Excel, html en de nieuwe Microsoft-formaten WordML en SpreadsheetML. Ook slaagt het programma erin om ingescand materiaal om te zetten naar pdf-documenten die zich laten doorzoeken op inhoud.

WEL OF NIET? 8/10

- ▲ Erg snel en accuraat, leest ook streepjescode
- ▼ Geen erg goede handschriftherkenning

www.irislink.com

Nuance OmniPage 16

Razendsnel omzetten



Ook OmniPage heeft zijn sporen verdiend in de OCR-wereld. Wij ontvingen de dure Professional Edition, maar die bevat wel twee extra pakketten: de documentenmanager ScanSoft PaperPort Standard 11 en een pdf-converteerder ScanSoft PDF Create.
Ook OmniPage herkent probleemloos teksten uit boeken, magazines, kranten, enzovoort.



Eerst gaat het programma via de wizard **Scannerinstellingen** kijken of jouw scantoestel in zijn database zit. Is dat zo, dan kan je meteen aan de slag; anders zal het programma aan de hand van een reeks tests de optimale instellingen kiezen.
Het programma werkt in drie weergavevormen. De klassieke interface lijkt op die van de vorige versie en is dus vooral voor gebruikers die bekend zijn met het programma. Geavanceerde gebruikers zijn wellicht meer gebaat bij de flexibele weergave. Die werkt met tabbladen, zodat je een goed overzicht hebt op de verschillende functies. Wil je zonder al te veel moeite tekst omzetten, dan gebruik je de Snelle conversie-weergave. OmniPage was duidelijk de snelste van de drie.

Archiveringsassistent

Je kan ingescande tekst markeren of zwart maken om hem te verbergen voor nieuwsgierige ogen. OmniPage beschikt ook over een functie om rechtstreeks foto's van de digitale camera te halen. In de optie **Juridisch document** kan je niet alleen nummeringen weglaten, maar is het ook mogelijk om een handtekening of stempel van een document te verwijderen.
Met de archiveringsassistent maak je werksets aan, zodat je bepaalde documenten op steeds dezelfde manier verwerkt. Daarvoor moet je je wel eerst identificeren aan de hand van een streepjescode. Wie veel formulieren moet verwerken, kan ervoor zorgen dat alleen de invulzones worden gelezen, waarna die gegevens worden weggeschreven als door komma's gescheiden tekstbestanden. Zo'n bestand kan je dan achteraf eenvoudig importeren in een database. Leggen we een workshop uit Clickx op de scanner en kiezen we de volautomatische functie, dan plaatst het programma de afbeeldingen, de tekst, de opmaak en zelfs (ongeveer) het juiste lettertype in een Word-bestand.

WEL OF NIET? 9/10

- ▲ Erg soepele workflow, snel
- ▼ Enkele snufjes die de prijs van de standaardversie optrekken, horen eerder thuis in een zakelijk pakket

www.nuance.be

ABBYY FineReader 9.0

Russisch taalwonder

Ook het Russische ABBYY timmert al jaren aan de weg van de tekstherkenning. Het bedrijf in Moskou is sinds 1989 gespecialiseerd in kunstmatige intelligentie, OCR en toegepaste linguïstiek. FineReader 9.0 herkent 179 talen, het meeste van de drie pakketten. Na de installatie vinden we in de programmamap niet alleen de toepassing, maar ook een map met **SNELLE TAKEN.** Hiermee kan je een foto of tekst onmiddellijk omzetten in een Word-document, pdf-bestanden converteren naar Word (of omgekeerd), een afbeelding scannen of een tabel vertalen naar Microsoft Excel. Tijdens onze tests had het programma wel problemen met het herkennen van klein gedrukte tekst. Gelukkig gaf de toepassing zelf aan dat de nauwkeurigheid aanzienlijk verbetert door de scanresolutie te verhogen van de normale 300 dpi naar 600 dpi of hoger. ABBYY FineReader 9.0 herkent ook automatisch de taal of verschillende talen waaruit een document is opgebouwd. Wil je de taalselectie toch in eigen handen houden, dan open je via **MEER TALEN** de **TAAL EDITOR.**



Screenshots lezen

ABBYY FineReader 9.0 bevat een nieuwe technologie die luistert naar de ronkende naam 'Adaptive Document Recognition Technology' (ADRTTM). Dankzij deze techniek zou de opmaak van een document behouden blijven, maar ons kon het niet overtuigen. Teksten en afbeeldingen converteren verliep uitstekend, maar bij de opmaak ging het behoorlijk fout.
Het is ook mogelijk om OCR toe te passen op beelden van een digitale camera. Het pakket komt trouwens met een handige functie, de ABBYY Screenshot Reader, die de tekst en afbeeldingen van schermafbeeldingen netjes omzet in Word of Excel. FineReader werkt erg prettig, omdat de interface opzettelijk heel eenvoudig is gehouden. Het resultaat in pdf is echter minder nauwkeurig dan bij de andere twee programma's. Bovendien is het ook iets trager.

WEL OF NIET? 9/10

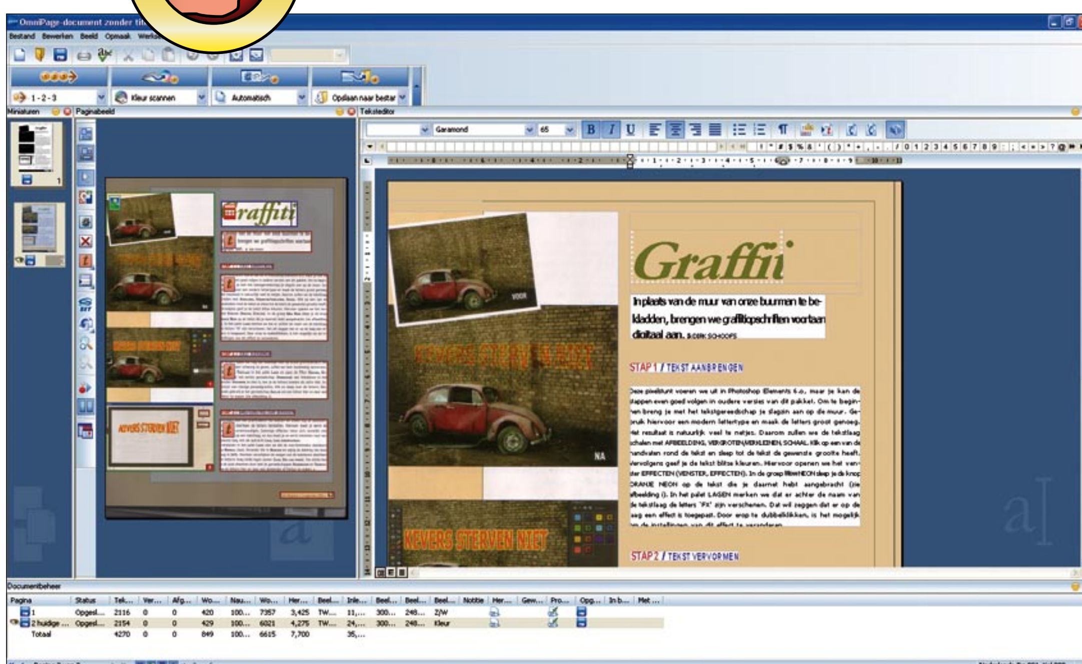
- ▲ Herkent maar liefst 179 talen, handige Screenshot Reader
- ▼ De beloofde opmaakherkenning stelt teleur

http://finereader.abby.com

De prijzen bij de producten zijn richtprijzen die door de producent zelf zijn doorgegeven. Toch blijken de meeste pakketten in de winkel een pak goedkoper. Readiris Pro was in een bepaalde shop zelfs twee derde goedkoper dan de aangegeven prijs.



Clickx keuze



OCR is heel wat geëvolueerd de laatste jaren. Zo is de intervalfunctie van Readiris Pro een nuttige vondst om grote hoeveelheden tekst efficiënt te verwerken, en bewijst ABBYY dat je niet alleen ingescande tekst kan omzetten, maar ook pakweg screenshots. Alle geteste pakketten zijn heel goede tekstherkenners. Op het vlak van snelheid springt **Nuance OmniPage 16** er uit. We zijn daarbij afgegaan op de mogelijkheden van de standaardversie. Het inscannen en omzetten gaat gewoon vlugger en bovendien kan je gelijksoortige documenten met één knop op dezelfde manier verwerken. Bovendien werkt dit pakket ook erg nauwkeurig, zodat we OmniPage 16 uitroepen tot Clickx keuze. ♦